

Query Expansion Using the UMLS[®] Metathesaurus[®]

Alan R. Aronson, Ph.D. and Thomas C. Rindflesch, Ph.D.
National Library of Medicine
Bethesda, MD 20894

ABSTRACT

Recent work has demonstrated the importance of query expansion for improving retrieval effectiveness when applying statistically-based systems to MEDLINE[®] citations. The research has suggested the use of retrieval feedback for enhancing the original text of users' queries. As an alternative method of query expansion, we propose the use of the MetaMap program for associating UMLS Metathesaurus concepts with the original query. Our experiments show that query expansion based on MetaMap compares favorably with retrieval feedback. We conclude that the optimal strategy would be to combine the two techniques.

INTRODUCTION

In a series of recent articles Srinivasan [1-4] explores information retrieval in MEDLINE, focusing on two areas of the retrieval process, indexing [1] and query expansion [2-4]. Her results with regard to query expansion can be summarized by quoting from [3]:

In brief, the best result obtained using queries expanded via retrieval feedback was a 11-AvgP (described later) score of 0.6018. This represents a statistically significant 16.4% improvement over baseline 11-AvgP score of 0.5169 obtained with unexpanded queries. (p. 354)

Srinivasan goes on to compare this result to other relevant work: "These papers show how the improvements in retrieval performance are superior to those obtained by the alternative methods..." [3:354] In contrasting her work to ours Srinivasan states: "In comparison with the approach of Aronson et al., the improvements achieved using retrieval feedback are much higher than the 4% improvement in precision that they report." [4:164] In comparing retrieval feedback to methods based on the UMLS Metathesaurus she says: "Although the metathesaurus offers an extremely rich vocabulary, it appears that further innovation and research are needed to make profitable use of this global medical thesaurus for information retrieval." [4:164] We consider this research to be interesting and significant and would like to examine

in greater detail its relationship to our work (Aronson et al., [5]), particularly with regard to the value of the UMLS Metathesaurus for information retrieval.

Several variables contribute to the final outcome in any IR experiment: the type of retrieval system, the indexing terms, the queries, and the test collection. If two systems are to be compared with regard to one of these variables the others must be kept constant (or controlled for). Srinivasan's work is based on documents which have human-assigned indexing terms associated with them. Her final results are due to the accumulated effect of this indexing as well as document feedback used for query expansion. The results reported in [5] are derived from documents without reference to human-assigned indexing terms and queries expanded through the use of the UMLS Metathesaurus. In order to felicitously compare our method of query expansion to Srinivasan's, both systems need to be based on the same type of indexing. Therefore, we would like to explore a comparison of our approach to Srinivasan's work, concentrating on query expansion, and keeping the indexing constant by using human-assigned MeSH[®] terms.

QUERY EXPANSION BASED ON RETRIEVAL FEEDBACK

Srinivasan [1] reports on a number of experiments which illustrate her overall approach. These experiments are based on a test collection produced by Hersh et al. [6] which comprises 75 queries and 2,344 MEDLINE citations. Each citation includes a title, an abstract, and MeSH indexing terms assigned by humans. The retrieval system used is SMART, a vector space model (Salton, [7]). All results are reported in terms of 11-point average precision.

Query expansion (Efthimiadis, [8]) is a technique for improving effectiveness in information retrieval systems. Research in automatic methods, which do not burden the user, is of particular importance. As a means of query expansion, Srinivasan employs retrieval feedback based on relevance feedback (Ide, [9]). In this technique salient terms from the top few relevant documents retrieved from an initial query are

used to produce a second, expanded query which performs better than the initial, unexpanded query. Relevance feedback is effective, but the user must choose the documents which contribute to the expanded query. In retrieval feedback the system automatically computes salient terms for inclusion in the expanded query, without input from the user. Tests in the context of the TREC experiments (Harmon, [10]) have demonstrated that retrieval feedback as a method of query expansion increases average precision, and Srinivasan's research supports those results.

In designing her experiments Srinivasan varies both the documents and the queries. Documents consist either of text only (MEDLINE title and abstract) or text along with the MeSH indexing terms assigned to that document. Queries are comprised of either text alone or are expanded with terms derived from retrieval feedback. The experiments which are relevant to this study are the following: a) both documents and queries are text-only (this is the baseline); b) documents contain text along with MeSH indexing terms and queries consist of text-only; and c) documents have both text and MeSH, and queries are expanded with retrieval feedback. It is significant to note that the query expansion process has access to the MeSH terms in the indexed documents and thus can add appropriate MeSH terms to the text of the query.

Srinivasan reports significant gains in average precision using her methodology. Average precision for retrieval of relevant documents in the baseline experiment, where both documents and queries consist of text-only, is 51.7%. Adding the MeSH terms to the documents but leaving the queries unexpanded increases average precision to 55.6%, which is a 7.5% increase over the baseline. Using expanded queries on MeSH-indexed text further increases average precision to 60.2%. This represents a 16.4% increase over the baseline average precision and an 8.3% increase over the experiment based on indexed text, but with unexpanded queries.

These experiments demonstrate that both human indexing as represented by the MeSH terms in the documents and retrieval feedback contribute to the overall improvement in average precision of 16.4% over the baseline figure. Apparently, each phenomenon contributes equally to the final result. Adding human-assigned MeSH terms (but leaving the queries unexpanded) adds 7.5% to the base score (51.7% to 55.6%). Similarly, adding retrieval feedback query expansion to MeSH-indexed documents adds 8.3% to the score with unexpanded queries (55.6% to 60.2%).

In a separate study [2] Srinivasan reports on query expansion based on a statistical thesaurus in addition to document feedback. The thesaurus contains statistical correlations between MeSH terms and text and is consulted during retrieval in order to add MeSH terms to the query. The thesaurus-based query expansion produces an increase in average precision over the baseline, but it is not as high as that achieved with feedback-based query expansion. The best result reported in [2] using a statistical thesaurus (and MeSH-indexed documents) is 56.8% average precision, which is a 9.9% increase over the baseline of 51.7%, but only 2.2% over the average precision attained with MeSH-indexed text and unexpanded queries. As a final step, Srinivasan combines her thesaurus-based query expansion with her feedback method and achieves a score of 60.5%. Since this represents a 17.1% increase over the baseline, it indicates that, although the results are not quite additive, retrieval feedback may be profitably combined with other methods of query expansion.

We felt it would be useful to explore the UMLS Meta-thesaurus-based query expansion discussed in [5] in the context of Srinivasan's research. In our earlier work we did not investigate the effects of applying our query expansion technique to MeSH-indexed text.

QUERY EXPANSION BASED ON METAMAP AND THE UMLS METATHESAURUS

In conducting experiments parallel to Srinivasan's, we used the test collection of Hersh et al. [6] and INQUERY (Callan et al., [11]), an inference network-based probabilistic retrieval model. This information retrieval system readily accommodates multiple representations for both documents and queries (Rajasehkar and Croft, [12]) and thus naturally supports documents either with or without MeSH indexing and queries in either original or expanded form.

INQUERY supports a variety of document processing techniques including the control of case, stopping, stemming and indexing. INQUERY's default behavior includes ignoring case, performing stopping, employing K-stemming and using a single index for all indexed fields. The experiments reported here rely on the default behavior except that Porter stemming is used instead of K-stemming. An example of a MEDLINE document in SGML format acceptable to INQUERY is

```
<DOC>
<DOCID>2102</DOCID>
<TITLE> Intravenous methylprednisolone efficacy in
status asthmaticus of childhood. </TITLE>
```

<TEXT> Forty-nine nonsteroid-dependent children hospitalized with status asthmaticus were randomized to receive IV placebo or methylprednisolone treatment (1 mg/kg every six hours). All patients received nebulized isoetharine inhalations and continuous IV aminophylline infusion. ... Findings of this study indicate that IV corticosteroid therapy is beneficial in treating pediatric status asthmaticus.
</TEXT>

<MESH>

Adolescence
Asthma/*DRUG THERAPY
Child
Clinical Trials
Double-Blind Method
Human
Intravenous Injections
Methylprednisolone/*ADMINISTRATION &
DOSAGE/THERAPEUTIC USE
Random Allocation
Status Asthmaticus/*DRUG THERAPY

</MESH>

</DOC>

The <DOCID> field contains the document's ID within the test collection, <TITLE> and <TEXT> consist of the title and abstract for the document, and <MESH> contains manually assigned MeSH terms for the document. Note that the <MESH> field is not present in documents used for baseline experiments.

As reported earlier in [5] and [13], the methodology used for expanding queries is MetaMap, a knowledge-based, linguistically rigorous program which maps biomedical text to concepts in the UMLS Metathesaurus. For example the text *is there evidence to support the use of inhaled steroids in COPD when the patient is on intravenous steroids* maps to the Metathesaurus concepts Supports, utilization, Inhaled, Steroids, Obstructive Lung Diseases, Patients and IV.

The expanded queries used for the INQUERY experiments reported here consist of the words from the original queries together with phrases and Metathesaurus concepts determined by MetaMap processing. MetaMap uses the SPECIALIST™ minimal commitment parser to segment the query text into simple noun phrases. It then performs intensive variant generation on each phrase and refers to a knowledge base of synonyms in order to discover the Metathesaurus concept(s) most related to the text. (In general MetaMap also uses knowledge bases of morphological variants and acronyms/abbreviations, but for this work no morphological variation was performed and only acronyms/abbreviations with unique expansions were allowed.)

The original queries are expanded using the phrases and concepts found by MetaMap. This is accomplished by forming a weighted sum of words, phrases and concepts using INQUERY's flexible query language. A series of experiments was conducted to determine optimum relative weights. One of the best weighting schemes is to assign words, phrases and concepts weights of 2, 1 and 5, respectively. Following is an example of an unexpanded query along with its expanded counterpart augmented with these weights (which appear immediately to the left of the #SUM operator).

Unexpanded query:

is there evidence to support the use of inhaled steroids in COPD when the patient is on intravenous steroids

Expanded query with weights:

```
#q34 =
#WSUM ( 1
  2 #SUM( is there evidence to support the use of
    inhaled steroids in COPD when the
    patient is on intravenous steroids )
  1 #SUM (
    #PHRASE (use)
    #PHRASE (inhaled steroids)
    #PHRASE (copd)
    #PHRASE (patient)
    #PHRASE (intravenous steroids))
  5 #SUM (
    #SUM ( Steroids )
    #SUM ( Obstructive Lung Diseases )
    #SUM ( Patients )
    #SUM ( utilization )
    #SUM ( Supports )
    #SUM ( Inhaled )
    #SUM ( IV )));
```

The #WSUM operator combines word, phrase and concept evidence with the chosen weights. In addition each phrase is delimited by the #PHRASE operator, and each concept is delimited by the #SUM operator. (INQUERY's #PHRASE operator can also be used for concepts, but the #SUM operator usually outperforms it.)

The following example illustrates the benefit provided by the query expansion technique just described. Relevant document 1431 with title *A comparison of metered dose inhalers with nebulizers from the delivery of ipratropium bromide in domiciliary practice* is ranked twelfth among the retrieved documents by the expanded query given above, but is not included in the top twenty retrieved documents with the original query. Scrutiny of the improved performance by the

expanded query reveals both the contribution of the query expansion method along with the value of MeSH indexing terms associated with the document.

The Metathesaurus terms for the expanded query do not obviously match the terms in the title any better than do the words from the plain text. However, note that one of the MeSH terms associated with this document is "Obstructive Lung Diseases/*DRUG THERAPY/PHYSIOPATHOLOGY." The MetaMap program determined that "Obstructive Lung Diseases" is a synonym of the expansion of the acronym COPD (Chronic Obstructive Pulmonary Disease) found in the title. This term is included in the expanded query and matches exactly to the main heading part of the MeSH term associated with the document. It is significant in this case that neither COPD or Obstructive Lung Diseases actually occurs in either the title or the abstract for this document. Thus query expansion alone would not have improved results for this query; it is only effective when it can be used with MeSH-indexed documents.

RESULTS

The results of applying the query expansion technique just described are summarized below (and in Figure 1), along with a comparison to the corresponding results obtained by Srinivasan. (All values are given in terms of 11-point average precision.)

A. Baseline: text only, no human-assigned MeSH terms, no query expansion

Srinivasan: 51.7

A&R: 52.3

B. Best results with human-assigned MeSH terms, no query expansion

Srinivasan: 55.6 (+7.5 over baseline)

A&R: 57.2 (+9.4 over baseline)

C. Best results with human-assigned MeSH terms, query expansion based on a thesaurus (A statistically-produced thesaurus for Srinivasan, UMLS Meta-thesaurus for A&R)

Srinivasan: 56.8 (+9.9 over baseline, +2.2 over no query expansion)

A&R: 59.7 (+14.1 over baseline, +4.4 over no query expansion)

D. Best results with human-assigned MeSH terms, query expansion based on a thesaurus combined with document feedback

Srinivasan: 60.2(+16.4 over baseline, +8.3 over no query expansion)

A&R: (not done)

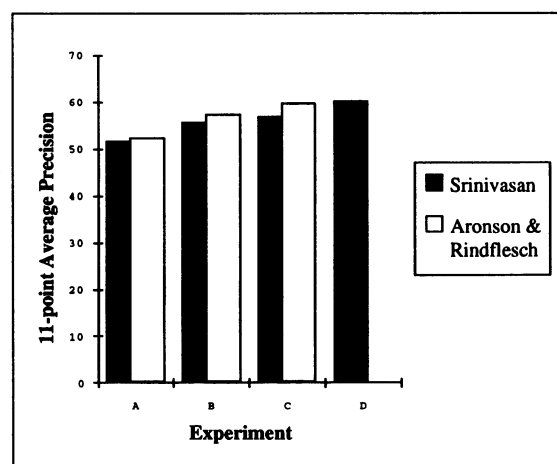


Figure 1. Comparison of Retrieval Results

DISCUSSION

For purposes of the current study, the salient results are those in C, where both systems take advantage of human-assigned MeSH terms and both systems use query expansion based on a thesaurus (a statistically-based thesaurus for Srinivasan, and the UMLS Meta-thesaurus for this study). The results, which are very comparable to Srinivasan's, seem to provide an accurate indication of the value of the Metathesaurus for query expansion. The somewhat better results from using the Metathesaurus may be significant or they may be due to the information retrieval system (INQUERY versus SMART).

We comment briefly concerning Srinivasan's final figure (in D). A number of factors may contribute to increases in average precision. Our final figure (59.7) using INQUERY, MeSH indexing, and UMLS Meta-thesaurus-based query expansion is not that far from Srinivasan's final figure (60.2) using SMART, MeSH indexing, and query expansion based on a statistically-produced thesaurus combined with document feedback.

We conclude that query expansion based on the UMLS Metathesaurus is an effective method of enhancing retrieval effectiveness and compares favorably to document feedback. We further feel that an

optimum system would also benefit from Srinivasan's insights by including document feedback along with UMLS-based query expansion. The two methods are not incompatible and can readily be combined (as Srinivasan herself does).

References

1. Srinivasan P. Optimal document-indexing vocabulary for MEDLINE. *Information Processing and Management*, 1996; 32(5):503-514.
2. Srinivasan P. Query expansion and MEDLINE. *Information Processing and Management*, 1996; 32(4): 431-443.
3. Srinivasan P. Retrieval feedback for query design in MEDLINE. A Comparison with Expert Network and LLSF Approaches. James J. Cimino (ed.) *Proceedings of the 1996 AMIA Annual Fall Symposium*, 1996:353-356.
4. Srinivasan P. Retrieval feedback in MEDLINE. *Journal of the American Medical Informatics Association*, 1996; 3(2):157-167.
5. Aronson AR, Rindflesch TC, and Browne AC. Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO 94*, 1994:197-216.
6. Hersh WR, Hickam DD, Haynes RB, and McKibbin KA. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association*, 1994; 1(1):51-60.
7. Salton G, Wong A, and Yang CS. A vector space model for automatic indexing. *Communications of the ACM*, 1975; 18(11):613-620.
8. Efthimiadis EN. Query expansion. In M.E. Williams (ed.) *Annual Review of Information Science and Technology*, 1996; 31:121-183. Published for the American Society for Information Retrieval by Information Today, Inc., Medford, NJ.
9. Ide E. New experiments in relevance feedback. In G. Salton (ed.) *The SMART Retrieval System-Experiments in Automatic Document Processing*, 1971:337-54. Englewood Cliffs, NJ: Prentice-Hall.
10. Harmon DK. Overview of the third Text REtrieval Conference (TREC-3). In D. K. Harmon (ed.) *Proceedings of the Third Text REtrieval Conference*, 1994:1-19.
11. Callan JP, Croft WB, and Harding S. The INQUERY retrieval system. *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, 1992:347-356.
12. Rajashekar TB, and Croft WB. Combining automatic and manual indexing representations in probabilistic retrieval. *Journal of the American Society for Information Science*, 1993;46(4):272-283.
13. Rindflesch TC, and Aronson AR. Ambiguity resolution while mapping free text to the UMLS Meta-thesaurus. In J. G. Ozbolt (ed.). *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care*, 1994:240-244.